ESANN'1997 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 16-17-18 April 1997, D-Facto public., ISBN 2-9600049-7-3, pp. 103-108

Sequential Hypotheses Tests for Modelling Neural Networks

Ulrich Anders, Olaf Korn

Centre for European Economic Research (ZEW)
P.O. Box 10 34 43
68034 Mannheim, Germany

Abstract.
This article examines how model specification in neural networks can be guided by statistical inference techniques. We develop a model selection strategy based on a sequence of statistical hypotheses tests, both LM–tests and Wald–tests. The strategy is evaluated in a simulation study and shown to be a very effective tool for network specification.

1. Introduction

Recently, there is growing interest in the modelling of nonlinear relationships. Unfortunately, for many applications theory does not guide the model building process by suggesting the relevant input variables or the correct functional form. This difficulty makes it attractive to consider an 'atheoretical' but flexible class of statistical models. Artificial neural networks are well suited for this purpose as it is known that they can approximate virtually any function up to an arbitrary degree of accuracy. This desired flexibility, however, makes the specification of an adequate network architecture even more difficult. Despite the huge amount of network theory and the importance of neural networks in applied work, there is still little experience with a statistical approach to model selection.

In this article we develop a model selection strategy for neural networks which consists of a sequence of statistical hypotheses tests. Taking a statistical perspective seems especially suited as it is just the lack of knowledge about an adequate functional form which calls for the application of neural networks. The specification strategy will be evaluated in a simulation study and found to be highly effective in identifying an adequate network architecture.

2. Neural Network Models

In this article we exclusively deal with so called 'multilayer perceptron networks'. Neural networks of this kind can be interpreted as a very flexible class of nonlinear regression models. The dependent variable y is predicted by the

ESANN'1997 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 16-17-18 April 1997, D-Facto public., ISBN 2-9600049-7-3, pp. 103-108

network output, which is a function $f(X, w)$ of the input data X and the network weights $w = (\beta', \gamma')'$. For a network of the type used in this paper the function takes the form

$$f(X, w) = \sum_{h=1}^{H} \beta_h g(\sum_{i=0}^{I} \gamma_{hi} x_i). \quad (1)$$

The variable x_0 is defined to be constant and set to $x_0 \equiv 1$. The scalars I and H denote the number of input and hidden units in the net and $g(\cdot)$ is a tangens hyperbolicus function attached to each hidden unit.

It is still the most unresolved question in the literature of neural networks what architecture is best to assume for a given problem. A desirable network architecture contains as little hidden units and connections as necessary for a good approximation of the true function, taking account of the tradeoff between estimation bias and variability due to estimation errors. Unfortunately, the form of the true function is seldomly known. It is therefore necessary to develop a methodology to select appropriate network models.

The usual approaches pursued in the network literature are *regularization*, *pruning*, and *stopped training*.[1] Although some of these methods may lead to satisfactory results, they comprise of a strong judgemental component, which makes the model building process difficult to reconstruct. This deficiency will be overcome by the strategy proposed in section 4. The next section briefly reviews statistical inference techniques for neural networks, which will serve as building blocks for the modelling strategy.

3. Hypotheses Testing in Neural Networks

Statistical inference in MLP–networks was developed by White (1989b). He showed that — if the parameters of a neural network are identified — they can be consistently estimated by maximum likelihood methods. Moreover, the parameter estimates of a network are asymptotically normally distributed. This knowledge in principle allows for the application of standard asymptotic hypotheses tests, such as Wald–tests or LM–tests.

Before parameter tests can be applied, however, it must be ensured that the parameters — apart from symmetric solutions — are uniquely identified. This is not the case when a network contains irrelevant hidden units. Since the β–weight of an irrelevant hidden unit would theoretically be zero, the γ–weights which lead into this hidden unit could take any value and are thus not identified. Similarly, if the γ-weights which lead into a hidden unit are all zero, the corresponding β-weight is equally not identified.

In other words, if we want to apply asymptotic normal distribution theory in neural network models, we must guarantee that the parameters are at least locally unique, i.e. there are no irrelevant hidden units in the network. Two techniques have been proposed in the literature yielding a χ^2-statistic

[1] Reed (1993) provides a survey of these methods.

ESANN'1997 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 16-17-18 April 1997, D-Facto public., ISBN 2-9600049-7-3, pp. 103-108

for testing the irrelevant hidden unit hypothesis and avoiding the identification problem. One technique was developed by White (1989a) and its properties investigated by Lee/White/Granger (1993). The other was devised by Teräsvirta/Lin/Granger (1993) and compared to the former.

White (1989a) suggests drawing the γ–weights of an additional hidden unit from a random distribution. This amounts to a random choice of the parameters in γ–space. The subsequent test on the β–weight of the additional hidden unit is carried out conditional to the random values of the γ–weights. Teräsvirta/Lin/Granger (1993) propose the application of a third order taylor expansion of the additional hidden unit, which equally leads to an avoidance of the identification problem. Both tests are performed in the fashion of a standard Lagrange multiplier test.[2]

When the network does not contain irrelevant hidden units, one can test for arbitrary parameter restrictions on the γ–weights by help of a Wald–test.

4. Model Selection Strategy

In the process of network architecture selection we have to guarantee model identification whenever inference techniques are used. Consequently, the strategy cannot adopt a top down approach which starts with a large (and probably over-parameterized) neural net. To obtain statistically valid results, the strategy begins with the smallest model possible and successively adds more complexity.

The strategy runs as follows: As a starting point all I input variables are combined with one hidden unit and the relevance of this unit is tested by the LM–test procedures of White (1989a) or Teräsvirta/Lin/Granger (1993). If the test fails to show significance, the whole procedure would stop; if the unit is relevant, it is included in the model. In this case the network is estimated and a further fully connected hidden unit tested for significance. The procedure continues until no further additional hidden unit shows relevance. After the number of hidden units is determined Wald–tests are applied in a top down approach to decide on the significance of single input connections. If there are insignificant connections, the one with the highest p-value is removed from the model and the reduced network retrained thereafter. This procedure is carried on until only significant connections remain in the model.

The proposed strategies ensure that the percentage of selected models which are overparameterized with respect to the number of hidden units is bounded by the sizes of the LM–tests.[3] In how far the procedure favours too small models depends on the power of the tests and will be investigated in the simulation.

[2] A detailed explanation of both procedures is given in Anders/Korn (1996).

[3] The test sizes may be different from the chosen significance levels in finite samples. Simulation results for the size of White's neural network test are given in Lee/White/Granger (1993)

ESANN'1997 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 16-17-18 April 1997, D-Facto public., ISBN 2-9600049-7-3, pp. 103-108

5. Simulation Study and Results

In the simulation study we consider three network models of different complexity. The models consist of three, five and seven variable inputs in addition to a constant input c and have three, four and five hidden units, respectively. In all networks there is one linear output unit. Thus the architecture of the networks can be expressed by 3c-3-1, 5c-4-1 and 7c-5-1. None of the networks is fully connected. Instead the connections are chosen due to the following rule: link the first input to all hidden units, the second input to all but the last hidden unit, the third input to all but the last two hidden units and so on.

The independent variables X are drawn from a standard normal distribution, while the β- and γ-weights are selected such that the outputs of the hidden units are as far uncorrelated as possible, in order to give the hidden units a high justification. Finally zero mean normal error terms are added to the conditional mean of y. The standard deviation σ_ε of the noise is chosen to be either ten or twenty percent of the conditional standard deviation $\sigma_{E[y|X]}$ of the network's output. The whole set of simulated data consists of 2000 observations, which we split into an in-sample training set and an out-of-sample set with 1000 data points each.

In the simulation we compare the out-of-sample mean squared prediction errors (MSPE) of the true model, a model with the true model structure but estimated weights and a model resulting from our selection strategy.[4] All tests employed in the strategy, LM–tests as well as Wald–tests, are carried out on a significance level of 5 percent. We repeated the experiment a hundred times, each time redrawing the in-sample random errors. The best out-of-sample performance which the model selection strategy can — apart from chance — achieve, is the out-of-sample MSPE of the true model. As we leave the out-of-sample noise unchanged, this value is taken as a benchmark.

The results of the simulation study are given in the following tables. Column 1 contains the model abbreviations. Column 2 reports how far the model with the true structure (TS) and estimated weights deteriorates from the true model (TM). This is measured by the relative differences in the out-of-sample MSPEs of the two models calculated as $(\mathrm{MSPE}_{\mathrm{TS}} - \mathrm{MSPE}_{\mathrm{TM}})/\,\mathrm{MSPE}_{\mathrm{TM}}$. The numbers given in the tables are the averages over the hundred replications of the simulation study.

Column 3 provides the corresponding results for the models chosen by the model selection strategy (MSS). Column 4 reveals the deterioration of MSPEs when moving from the TS to the MSS–models. Column 5 finally shows how often the model selection strategy found a number of hidden units (NH) that was smaller, equal or larger than the true number of hidden units.

The simulation results for the model selection strategy employing the LM–test of Teräsvirta/Lin/Granger (1993) are encouraging. In the case of the smaller models (3c-3-1 and 5c-4-1) the out-of-sample performance of the selec-

[4] To reduce the problems of converge to a local maximum of the likelihood function we ran the model selection strategy several times with different starting values and took the model yielding the smallest in-sample mean squared error.

ESANN'1997 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 16-17-18 April 1997, D-Facto public., ISBN 2-9600049-7-3, pp. 103-108

tion strategy is less than 4% worse than that of the true model (column 3). This is a particularly small value as even the knowledge of the true network architecure cannot reduce the MSPE much further. The difference between the models of the true structure and the ones selected by sequential testing is less than 2% on average. In some cases the model selection strategy even improved the out-of-sample performance of the true model.

Models	TS vs. TM	MSS vs. TM	TS vs. MSS	NH <, =, >
3c-3-1	1.12%	1.43%	0.30%	00 , 74 , 26
5c-4-1	2.07%	2.88%	0.79%	00 , 85 , 15
7c-5-1	3.97%	5.97%	1.92%	00 , 49 , 51

Table 1: $\sigma_\varepsilon = 0.1\sigma_{E[y|X]}$, using LM–test of Teräsvirta/Lin/Granger.

Models	TS vs. TM	MSS vs. TM	TS vs. MSS	NH <, =, >
3c-3-1	1.20%	1.55%	0.34%	04 , 88 , 08
5c-4-1	2.37%	3.67%	1.27%	02 , 87 , 11
7c-5-1	3.40%	13.53%	9.80%	55 , 43 , 02

Table 2: $\sigma_\varepsilon = 0.2\sigma_{E[y|X]}$, using LM–test of Teräsvirta/Lin/Granger.

Models	TS vs. TM	MSS vs. TM	TS vs. MSS	NH <, =, >
3c-3-1	1.00%	3.54%	2.52%	20 , 75 , 05
5c-4-1	2.18%	13.83%	11.40%	95 , 05 , 00
7c-5-1	3.66%	56.96%	51.40%	89 , 11 , 00

Table 3: $\sigma_\varepsilon = 0.1\sigma_{E[y|X]}$, using LM–test of White.

Models	TS vs. TM	MSS vs. TM	TS vs. MSS	NH <, =, >
3c-3-1	0.98%	3.05%	2.05%	60 , 40 , 00
5c-4-1	2.53%	7.02%	4.39%	100 , 00 , 00
7c-5-1	3.72%	24.90%	20.43%	100 , 00 , 00

Table 4: $\sigma_\varepsilon = 0.2\sigma_{E[y|X]}$, using LM–test of White.

The true number of hidden units is found by the model selection strategy in at least 74% of all replications. When the true number of hidden units is not found the strategy seems to overestimate the number of hidden units. The percentage of models with too many hidden units clearly exeeds the chosen 5 percent significance level of the LM–tests. It should be noted, however, that even if we end up with too many hidden units, the resulting model is not generally overparameterized. Since in the top down step of the specification strategy several γ-weights might have been removed, the overall number of parameters may even be smaller than for the true model.

If the true model becomes larger (7c-5-1) the out-of-sample MSPE grows, but it is still less than 10% higher than for the true structure. It is interesting to note that the selection strategy tends to underestimate the number of hidden units for the large model with $\sigma_\varepsilon = 0.2\sigma_{E[y|X]}$. This may indicate that the power of the tests is not sufficient for the given noise level and the available

number of one thousand data points in the training set.

The results of the model selection strategy employing the LM–test of White (1989a) are shown in tables 3 and 4. The performance of the model selection strategy is still reasonable but clearly inferior compared with the results in tables 1 and 2. The worse performance of the selection strategy based on the White test stems from an underfitting of the true model, as many of the chosen specifications contain less than the true number of hidden units. The LM–test proposed by White (1989a) seems to have less power than the method of Teräsvirta/Lin/Granger (1993).

6. Conclusion

In this article we introduced a systematic approach to model selection in neural networks. Our model selection strategy consists of a sequence of hypotheses tests, where LM–tests are followed by Wald–tests. In order to avoid the non-identification of neural networks we employed two alternative methods due to White (1989a) and Teräsvirta/Lin/Granger (1993).

The model selection strategy was evaluated in a simulation study. It turned out that the selection strategy based on the LM–test of Teräsvirta/Lin/Granger (1993) is clearly superior. When this test is employed, the strategy leads to models with a very good out-of-sample performance, being not much worse than the true model. In most cases the strategy identifies the correct number of hidden units and therefore achieves a good balance between under- and overfitting. The overall results of the simulation provide evidence that the proposed strategy is a powerful tool for neural network model building. This is promising for applications with real data sets.

References

[1] Anders U., Korn O. (1996): *Model Selection in Neural Networks.* ZEW Discussion Paper 96-21.

[2] Lee T.-H., White H., Granger C.W.J. (1993): *Testing for Neglected Nonlinearity in Time Series Models.* Journal of Econometrics, 56, 269–290.

[3] Reed R. (1993): *Pruning Algorithms — A Survey.* IEEE Transactions on Neural Networks, 4, 740–747.

[4] Teräsvirta T., Lin C.-F., Granger C.W. (1993): *Power of the Neural Network Linearity Test.* Journal of Time Series Analysis, 14 (2), 209–220.

[5] White H. (1989a): *An Additional Hidden Unit Test for Neglected Nonlinearity in Multilayer Feedforward Networks.* Proceedings of the International Joint Conference on Neural Networks. SOS Printing, II, 451–455.

[6] White H. (1989b): *Learning in Neural Networks: A Statistical Perspective.* Neural Computation, 1, 425–464.